15

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 06-162167
(43)Date of publication of application : 10.06.1994

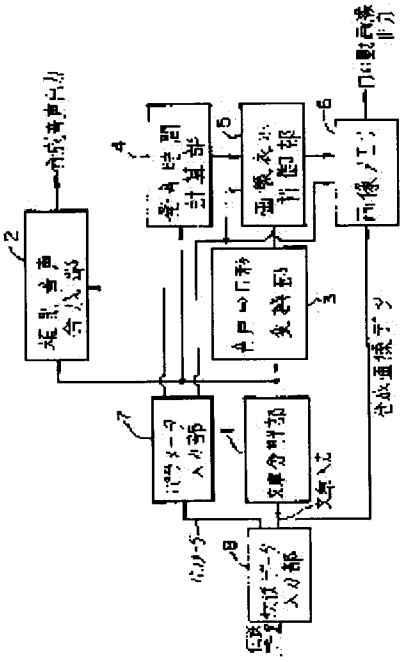
(51)Int.Cl. G06F 15/62
G06F 3/153
G06F 3/16
G10L 3/00

(21)Application number : 04-335526
(22)Date of filing : 20.11.1992

(71)Applicant : FUJITSU LTD
(72)Inventor : NAKAGAWA AKIRA
MORIMATSU EIJI
MATSUDA KIICHI

(54) COMPOSITE IMAGE DISPLAY SYSTEM

(57)Abstract:
PURPOSE: To make a display exactly as a sending-side person intends by embedding parameters of a composite image that a person who composes data desires on an expressing side in the composite data on a face image and using the values em bedded in the image data as initial values of a system on a display side.
CONSTITUTION: Mapping to a face model is performed on the basis of the original image of the desired face image to be displayed on a reception side, and parameters of respective mouth shapes are used to generate the composite image data, in which the parameters regarding an impression to be given to the opposite reception side are embedded. Document information is generated separately and the both are sent as transmitted data to the reception side. The data are inputted to a transmitted data input part 8; and document information is sent to a document decomposition part 1, the composite image data is sent to an image memory 6, and various parameters are sent to a parameter input part 7. The parameter input part 7 once receiving the parameters checks and sends them to an image display control part 5 and the image memory 6. On the reception side, the information is displayed exactly as the sending-side person intends.



* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1]In an image composing display system which generates synthetic video of a person's face in which a mouth moves according to synthesized speech and this synthesized speech corresponding to it from arbitrary text data, An image composing display system constituting so that various parameters for deciding a generation mode of synthesized speech and synthetic video by the side of a display may be added to this image composing and the display side may be passed, when creating image composing of a face to the creation side of text data.
[Claim 2]The image composing display system according to claim 1 which is a parameter including display magnification at the time of these various parameters displaying vocal quality of synthesized speech, and synthetic video, and a display position.
[Claim 3]An image composing display system comprising:
A transmission data input means which divides received transmission data into composited dynamic image data, text data, and various parameters.
A voice synthesis means which generates and outputs synthesized speech based on these text data.
An image memory which files composited dynamic image data separated by this transmission data input means.
A conversion method changed into a series of mouth form numerals showing a motion of a series of mouth type when the text data are uttered for these text data, A pronunciation time calculation means which calculates pronunciation time of each syllable of synthesized speech outputted from this voice synthesis means based on these text data, and presumes timing of a break of each sound, A display control means which performs control which switches a display image to a mouth form picture corresponding to mouth form numerals from this conversion method in timing of a break of each syllable presumed by this pronunciation time calculation means, and a parameter input means to send various parameters separated by this transmission data input means to a corresponding internal circuit.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Industrial Application]This invention relates to an image composing display system applicable to AV (audio video) E-mail etc. which can tell the other party a message by the synthetic video and synthesized speech of the face with which the sending person has talked just like TV telephone only by sending text (text) data, and especially, It is related with the image composing display system which can tell impressions which are the creation sides of a text etc. and were meant, such as vocal quality and face expression, to the display side.
[0002]
[Description of the Prior Art]The art of generating the synthesized speech corresponding to it freely, and pronouncing it from arbitrary text (text) information is called rule voice synthesis, and the rule voice synthesizer for realizing this is already made. This rule speech synthesis technique is applied in various fields in order to raise the interface of human being and machinery. The art which generates the video of a person including a motion of a mouth when it is spoken from arbitrary text data in analyzing the text data is developed like audio composition in recent years, By combining this with above-mentioned speech synthesis technique, a more natural interface is realizable.
[0003]For example, by preparing beforehand data files, such as a transmitting mail person's face picture, for the receiver, if the synthetic art of this sound and face video is applied to an E-mail, In the former, the message of a rich expression of the video of a face in which the transmitting mail person has talked appearing to the E-mail with which the text was only displayed on the screen of a receiver, and reading out by synthesized speech can be told to an addressee.
[0004]The example of composition of the sound and video output unit which compounds and outputs a sound and face video based on such a text is shown in drawing 4. In drawing 4, 1 is a text decomposition part into which text (text) information is inputted, and this text decomposition part 1 analyzes the inputted text data, generates the sound control data for voice response, and outputs it to the rule speech synthesis section 2, and the sound / mouth form converter 3. as text data -- "now" -- when a text is inputted, this is decomposed and outputted to the phoneme data which consists of the vowel and consonant of "T, A, D, A, I, M, A."
[0005]The rule speech synthesis section 2 is a device which generates and outputs the synthesized speech which reads out the text based on the phoneme data about arbitrary texts.
[0006]A sound / mouth form converter 3 is the devices for changing into the series of the mouth form numerals for expressing a motion of a series of mouths at the time of pronouncing the text for the phoneme data about arbitrary texts. As mouth form numerals, there are seven kinds, A (vowel A), I (vowel I), U (vowel U), E (vowel E), O (vowel O), S (consonant), and C (closed mouth), and the picture of the mouth type at the time of pronouncing them corresponding to each mouth form numerals is prepared beforehand. as text data -- the above-mentioned -- "now", when a text is inputted, Based on the phoneme data "TADAIMA" of the text, "T" -> mouth type numerals S. "A" -> mouth type "numerals A and D" -> the mouth form "numerals S and I" -> mouth type "numerals I and M" -> mouth type "numerals C and A" -> mouth type numerals A are assigned, respectively, and they are outputted to the image display

controller 5 as a series of mouth form numerals.
[0007]Composite image data is filed in the image memory 6. As this composite image data, the data of the shoulder top picture for one frame of a speaker and seven kinds of mouth region images corresponding to seven kinds of above-mentioned mouth form numerals compounded based on it is gathered, and it is considered as one file.
[0008]The pronunciation time calculation part 4 calculates time until each syllable at the time of compounding a sound using the completely same algorithm as the rule speech synthesis section 2 based on the sound control data from the text decomposition part 1 is pronounced, respectively. That is, when the pronunciation output of it is synthesized voice from it and carried out by the rule speech synthesis section 2 to the inputted text, the timing of the break of each syllable which constitutes the text with the head of a text as the starting point is presumed, respectively, and the result is outputted to the image display controller 5.
[0009]The image display controller 5 performs image display control so that the mouth form picture corresponding to the mouth form numerals of the applicable syllable may be chosen from the image memory 6 and may be outputted, when the pronouncing timing of each syllable comes based on the timing signal from the pronunciation time calculation part 4. That is, synchronous control is performed so that the synchronization with synthesized speech and face video can be taken, so that a motion of a speaker's mouth displayed on a screen to the sound pronounced by the rule speech synthesis section 2 may be in agreement that is,.
[0010]The vocal quality of the sound which compounds the parameter inputting part 7 by the rule speech synthesis section 2, the display place on the screen of face video, It is a portion which inputs various parameters, such as display magnification, using a keyboard etc., and the parameter about synthesized speech is passed to the rule speech synthesis section 2, and the parameter about face video is passed to the image display controller 5 and the image memory 6.
[0011]Operation of the device constituted in this way is explained. If text data are inputted, the text data will be analyzed by the text decomposition part 1, phoneme data will collect, the rule speech synthesis section 2 will be passed, and a pronunciation output will be carried out by synthesized speech. In parallel to this pronunciation operation, phoneme data is changed into the series of mouth form numerals by the sound / mouth form converter 3. In the pronunciation time calculation part 4, the time of the break of each syllable is presumed from phoneme data, and this temporal data is passed to the image display controller 5. The timing of mouth form numerals is doubled with the pronouncing timing of each syllable in the image display controller 5, The face dynamic image data corresponding to the mouth form numerals which were able to be found in the sound / mouth form converter 3 from the inside of the picture of each mouth form numerals developed on the image memory 6 is transmitted to VRAM, and displays a speaker's face video on the screen of a display via this VRAM. Text data will be given to an addressee by this as a message by the face video of the speaker who had the timing of a motion of a mouth in the synthesized speech which actually pronounced it, and its synthesized speech.
[0012]The device of this drawing 4 is realizable as a small and economical system by using a certain small voice synthesis unit for the rule speech synthesis section 2 from the former, and using a personal computer etc. for the other portion.
[0013]
[Problem(s) to be Solved by the Invention]When realizing this sound and face video output unit on a personal computer, generally what is switched and displayed according to the text which creates image composing beforehand, and into which those pictures were inputted as mentioned above is performed for throughput reduction. In generating synthesized speech and face video in these devices, what was beforehand set as the system which displays parameters, such as vocal quality, a display place of the picture on a screen, and display magnification, as an initial value (what was beforehand inputted by the parameter inputting part 7) is used.
[0014]Thus, although the generation mode of face video is beforehand set to the vocal quality of synthesized speech by a receiver with the conventional device, unnatural sensibility will be given to those who see it when the person and vocal quality of the face picture which have been ******** registered do not balance to the message.
[0015]So that it may be represented, when this device is used for an E-mail etc., When those

who actually display and look at text data, the person who made image composing, and its text data with a sound and video differ, in the size of vocal quality and a picture for which the person who made text data and image composing wishes. In a receiver, pronunciation and the impression which image display is not always carried out and is completely different from an intention of the person of the delivery side as a result may be given to the person of a receiver.
[0016]That is, it will be decided with the parameter which the impression which the person of the display side receives from the vocal quality which hits telling a message with a sound and video in the conventional device, the looks of a face, etc. is a display side, and was set beforehand, It was not able to tell exactly the person of the display-impression expression which person of the creation side of information meant side.
[0017]This invention is made in view of this problem, and it is a display side, and in displaying synthesized speech or face image composing based on text data, the place made into the purpose is to make it a display as the person of the creation side of the text meant attained.
[0018]
[Means for Solving the Problem]Drawing 1 is a principle explanatory view concerning this invention. In an image composing display system which generates synthetic video of a person's face in which a mouth moves as one gestalt according to synthesized speech and this synthesized speech corresponding to it from arbitrary text data in this invention, When creating image composing of a face to the creation side of text data, an image composing display system constituting so that various parameters for deciding a generation mode of synthesized speech and synthetic video by the side of a display may be added to image composing and the display side may be passed is provided.
[0019]The above-mentioned various parameters can be made into a parameter including display magnification at the time of displaying vocal quality of synthesized speech, and synthetic video, and a display position.
[0020]A transmission data input means which divides received transmission data into composited dynamic image data, text data, and various parameters as other gestalten in this invention, A voice synthesis means which generates and outputs synthesized speech based on text data, and an image memory which files composited dynamic image data separated by a transmission data input means, A conversion method changed into a series of mouth form numerals showing a motion of a series of mouth type when the text data are uttered for text data, A pronunciation time calculation means which calculates pronunciation time of each syllable of synthesized speech outputted from this voice synthesis means based on text data, and presumes timing of a break of each sound, A display control means which performs control which switches a display image to a mouth form picture corresponding to mouth form numerals from this conversion method in timing of a break of each syllable presumed by a pronunciation time calculation means, An image composing display system provided with a parameter input means to send various parameters separated by a transmission data input means to a corresponding internal circuit is provided.
[0021]
[Function]In the image composing display type of this invention, in the transmitting side, when a face picture required for a display is combined, the display magnification of image composing, the display position, the vocal quality of synthesized speech, and other parameters which the person who is a display side and compounded to the same data wishes are embedded. In the display side, the value embedded at the image data is used as an initial value of a system. Synthesized speech and image composing are generable by the display system side as the person who combined the face picture meant by this.
[0022]In the image composing display system of other gestalten of this invention, The transmission data received by the transmission data input means Composited dynamic image data, text data, Separate into various parameters, analyze text data by a text decomposing means, and sound control data is generated, Based on this sound control data, synthesized speech is generated and outputted by a voice synthesis means, File the received composited dynamic image data in an image memory, and sound control data is changed into the series of mouth form numerals by a conversion method, Calculate the pronunciation time of each syllable

pronounced by a voice synthesis means based on sound control data by a pronunciation time calculation means, respectively, and the timing of the break of each syllable is presumed, It controls to read the mouth form picture of the syllable from an image memory according to the timing signal of each syllable by an image display control means, The various parameters which received are sent to an internal circuit corresponding by a parameter input means, and it is made to become synthesized speech and the thing in which the person of the text creation side meant the generation mode of image composing.

[0023]

[Example]Hereafter, the example of this invention is described with reference to drawings. A sound and a face video output unit according [drawing 2] to the image composing display system as one example of this invention are shown. In drawing 2, the text decomposition part 1, the rule speech synthesis section 2, the sound / mouth form converter 3, the pronunciation time calculation part 4, the image display controller 5, and the image memory 6 are the same as what was explained by the above-mentioned conventional example.

[0024]To the transmission data sent from the transmitting side, as a point of difference with a device conventionally. The parameter of the display magnification on composite image data, such as a face picture the person of the transmitting side other than original text data expects what is compounded and displayed by a receiver, and a mouth form picture, and the screen which the person wishes to have further, a display position, vocal quality, and others is embedded at composite image data.

[0025]The concept of the processing for embedding these parameters to transmission data in the transmitting side is shown in drawing 3. Mapping to a face model is performed based on the original image of the face picture which wishes the display by a receiver, composite image data is created using the parameter of each mouth type, and the parameter concerning the impression given to this at the other display magnification, display position, vocal quality, and receptacle side is embedded. Text data are created apart from this, and both sides are used as transmission data and seen off in a receiver. In this case, once it sends the composite image data in which the parameter was embedded, the rest should just repeat and send text data.

[0026]In a receiver, this transmission data is inputted into the transmission data input part 8, it separates into text data, composite image data, and various parameters, and send text data to the text decomposition part 1, composite image data is sent to the image memory 6, and various parameters are sent to the parameter inputting part 7 here, respectively.

[0027]If the parameter inputting part 7 receives these various parameters, these various parameters will be investigated and parameters, such as vocal quality about voice synthesis, will send the parameter about the picture of the display magnification of a picture, a display position, etc. to the image display controller 5 and the image memory 6 at the rule speech synthesis section 2, respectively.

[0028]With constituting in this way, the parameter of the face picture for which the person of the transmitting side wished, display magnification and a display position, vocal quality, and others can be used at a receiver as a parameter embedded as the face picture which should be displayed, and an initial value of a system. Therefore, a message can be displayed on the display system of a receiver by the sound and picture as an intention of the person of the transmitting side.

[0029]In operation of this invention, various modification gestalten are possible. For example, although the above-mentioned example explained the case where seven kinds of pictures were used as a mouth form, of course, this invention is not restricted to this, and in order to compound a motion of the mouth nearer to nature, it may increase the kind of picture of this mouth type further. Although the motion of a mouth field was taken up in the above-mentioned example as a motion portion of the face picture which is a display side and is combined, if it is not restricted to this, and a mouth moves, for example, and it is alike, in addition is made to change a motion of eyes etc. according to a text, AV message of a richer expression can be sent to the receptacle side.

[0030]Although the above-mentioned example explained the case where this invention was applied to AV E-mail, It is also possible for this invention not to be restricted to this and to

apply to a sound and a face video output unit simple substance, and, Or if recognition of the phoneme of an utterance sound is attained, for example by speech recognition technology in real time, it is also possible to apply to service of the false TV phone that the expression of a speaker's face can also be displayed on the addressee side with video only by making the usual phone call etc.

[0031]

[Effect of the Invention]As explained above, according to this invention, in displaying synthesized speech or face image composing based on text data by a receiver, a display as the person of the delivery side of the text meant is attained.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1]It is a principle explanatory view concerning this invention.
[Drawing 2]It is a figure showing the sound and video output unit by the image composing display system as one example of this invention.
[Drawing 3]It is a figure explaining the processing concept by the side of delivery by an example system.
[Drawing 4]It is a figure showing the conventional sound and video output unit.
[Description of Notations]
1 Text decomposition part
2 Rule speech synthesis section
3 A sound / mouth form converter
4 Pronunciation time calculation part
5 Image display controller
6 Image memory
7 Parameter inputting part
8 Transmission data input part

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号

特開平6－162167

(43)公開日　平成６年(1994)６月10日

(51)Int.Cl.[5]		識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ	15/62	３４０	8125－5L		
	3/153	３２０　Ｌ	7165－5B		
	3/16	３３０　Ｃ	7165－5B		
Ｇ１０Ｌ	3/00	Ｓ	8946－5H		

審査請求　未請求　請求項の数３(全　７　頁)

(21)出願番号　特願平4－335526

(22)出願日　平成４年(1992)11月20日

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中1015番地

(72)発明者　中川　章
神奈川県川崎市中原区上小田中1015番地
富士通株式会社内

(72)発明者　森松　映史
神奈川県川崎市中原区上小田中1015番地
富士通株式会社内

(72)発明者　松田　喜一
神奈川県川崎市中原区上小田中1015番地
富士通株式会社内

(74)代理人　弁理士　小林　隆夫

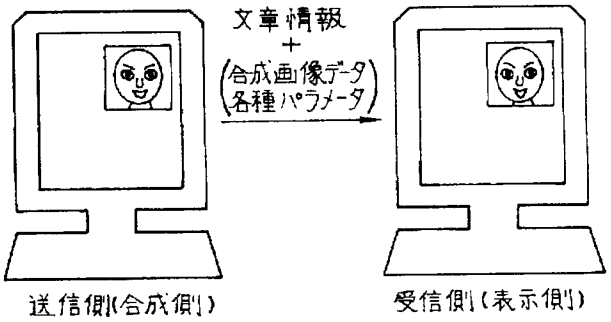
(54)【発明の名称】　合成画像表示システム

(57)【要約】

【目的】　文章（テキスト）データを送るだけであたかもＴＶ電話のように送信者が喋っている顔の合成動画像と合成音声で相手側にメッセージを伝えることができるＡＶ（オーディオ・ビデオ）電子メール等に適用できる合成画像表示システムに係り、特に、文章作成側で意図した声質や顔表情等の印象を的確に表示側に反映させることができる合成画像表示システムに関するものであり、表示側で文章情報に基づいて合成音声あるいは顔合成画像を表示するにあたり、その文章の作成側の人が意図した通りの表示が可能となるようにすることを目的とする。

【構成】　任意の文章情報からそれに対応する合成音声および合成音声に合わせて口が動く人物の顔の合成動画像を生成する合成画像表示システムにおいて、文章情報の作成側において顔の合成画像を作成する際に、表示側における合成音声と合成動画像の生成態様を決めるための各種パラメータを合成画像データに付加して表示側に渡すように構成されたことを特徴とする。

本発明に係る原理説明図



1

【特許請求の範囲】

【請求項１】　任意の文章情報からそれに対応する合成音声および該合成音声に合わせて口が動く人物の顔の合成動画像を生成する合成画像表示システムにおいて、
文章情報の作成側において顔の合成画像を作成する際に、表示側における合成音声と合成動画像の生成態様を決めるための各種パラメータを該合成画像に付加して表示側に渡すように構成されたことを特徴とする合成画像表示システム。

【請求項２】　該各種パラメータは合成音声の声質、合成動画像を表示する際の表示倍率、表示位置を含むパラメータである請求項１記載の合成画像表示システム。

【請求項３】　受信した伝送データを合成動画像データ、文章情報、各種パラメータに分離する伝送データ入力手段と、
該文章情報に基づいて合成音声を生成し出力する音声合成手段と、
該伝送データ入力手段で分離された合成動画像データをファイリングする画像メモリと、
該文章情報をその文章情報を発声したときの一連の口形の動きを表す口形符号の系列に変換する変換手段と、
該文章情報に基づいて該音声合成手段から出力される合成音声の各音節の発音時間を計算して各音声の切れ目のタイミングを推定する発音時間計算手段と、
該発音時間計算手段で推定した各音節の切れ目のタイミングで表示画像を該変換手段からの口形符号に対応した口形画像に切り換える制御を行う表示制御手段と、
該伝送データ入力手段で分離された各種パラメータを対応する内部回路に送るパラメータ入力手段とを備えた合成画像表示システム。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、文章（テキスト）データを送るだけであたかもＴＶ電話のように送信者が喋っている顔の合成動画像と合成音声で相手側にメッセージを伝えることができるＡＶ（オーディオ・ビデオ）電子メール等に適用できる合成画像表示システムに係り、特に、文章等の作成側で意図した声質や顔表情等の印象を表示側に伝えることができる合成画像表示システムに関するものである。

【０００２】

【従来の技術】任意の文章（テキスト）情報からそれに対応した合成音声を自由に生成し発音する技術は、規則音声合成と呼ばれ、これを実現するための規則音声合成装置が既に作られている。この規則音声合成技術は人間と機械とのインターフェースを向上させるために様々な分野で応用されている。また、近年、音声の合成と同様に、任意の文章情報からそれを喋ったときの口の動きを含む人物の動画像をその文章情報を解析することで生成する技術が開発されており、これを上述の音声合成技術

2

と組み合わせることによって、より自然なインターフェースを実現することができる。

【０００３】例えば、かかる音声と顔動画像の合成技術を電子メールに適用すると、受信側にメール送信者の顔画像などのデータファイルを予め用意しておくことにより、従来では受信側の画面上に文章が表示されるだけであった電子メールに対して、メール送信者が喋っている顔の動画像が現れて合成音声で読み上げるといった表現豊かなメッセージを受信者に伝えることができる。

【０００４】このような文章に基づいて音声および顔動画像を合成し出力する音声・動画像出力装置の構成例を図４に示す。図４において、１は文章（テキスト）情報が入力される文章分解部であり、この文章分解部１は入力された文章情報を解析して音声出力用の発音制御データを生成し規則音声合成部２と音声／口形変換部３に出力する。例えば、文章情報として「ただいま」の文章が入力された場合、これを「Ｔ，Ａ，Ｄ，Ａ，Ｉ，Ｍ，Ａ」の母音と子音からなる音素データに分解して出力する。

【０００５】規則音声合成部２は任意の文章についての音素データに基づいてその文章を読み上げる合成音声を生成し出力する装置である。

【０００６】音声／口形変換部３は、任意の文章についての音素データをその文章を発音する際の一連の口の動きを表すための口形符号の系列に変換するための装置である。口形符号としては例えば、Ａ（母音のア）、Ｉ（母音のイ）、Ｕ（母音のウ）、Ｅ（母音のエ）、Ｏ（母音のオ）、Ｓ（子音）、Ｃ（閉じた口）の７種類があり、それぞれの口形符号に対応してそれらを発音する際の口形の画像が予め用意される。例えば、文章情報として前述の「ただいま」の文章が入力された場合、その文章の音素データ「ＴＡＤＡＩＭＡ」に基づいて、「Ｔ」→口形符号Ｓ、「Ａ」→口形符号Ａ、「Ｄ」→口形符号Ｓ、「Ｉ」→口形符号Ｉ、「Ｍ」→口形符号Ｃ、「Ａ」→口形符号Ａ、をそれぞれ割り当てて、それらを口形符号の系列として画像表示制御部５に出力する。

【０００７】画像メモリ６には合成画像データがファイリングされている。この合成画像データとしては、話者の１フレーム分の肩上画像と、それを基に合成した前述の７種類の口形符号に対応した７種類の口領域画像のデータとを纏めて一つのファイルとしている。

【０００８】発音時間計算部４は文章分解部１からの発音制御データに基づいて規則音声合成部２と全く同じアルゴリズムを用いて音声を合成する際の各音節が発音されるまでの時間をそれぞれ計算する。つまり、入力された文章に対してそれが規則音声合成部２で音声合成されて発音出力される際に、文章の先頭を起点にしてその文章を構成する各音節の切れ目のタイミングをそれぞれ推定してその結果を画像表示制御部５に出力する。

【０００９】画像表示制御部５は発音時間計算部４からのタイミング信号に基づいて、各音節の発音タイミングが到来したときにその該当する音節の口形符号に対応する口形画像が画像メモリ６から選択されて出力されるよう画像表示制御を行う。すなわち、規則音声合成部２で発音される音声に対して画面に表示される話者の口の動きが一致するよう、つまり合成音声と顔動画像との同期がとれるように同期制御を行うものである。

【００１０】パラメータ入力部７は規則音声合成部２で合成する音声の声質、顔動画像の画面上での表示場所、表示倍率等の各種パラメータをキーボード等を用いて入力する部分であり、合成音声に関するパラメータは規則音声合成部２に渡され、また顔動画像に関するパラメータは画像表示制御部５と画像メモリ６に渡される。

【００１１】このように構成した装置の動作を説明する。文章情報が入力されると、文章分解部１でその文章情報が解析されて音素データがまとめて規則音声合成部２に渡されて合成音声により発音出力される。この発音動作に並行して、音素データが音声／口形変換部３で口形符号の系列に変換される。また発音時間計算部４では音素データから各音節の切れ目の時間が推定され、この時間データが画像表示制御部５に渡される。画像表示制御部５では各音節の発音タイミングに口形符号のタイミングを合わせて、画像メモリ６上に展開された各口形符号の画像のうちから音声／口形変換部３で求まった口形符号に対応した顔動画像データがＶＲＡＭに転送されるようにし、このＶＲＡＭを介して表示装置の画面上に話者の顔動画像を表示する。これにより文章情報は、それを実際に発音した合成音声とその合成音声に口の動きのタイミングがあった話者の顔動画像とによるメッセージとして受信者に伝えられることになる。

【００１２】この図４の装置は、規則音声合成部２に従来からある小型の音声合成ユニットを利用し、それ以外の部分にはパーソナルコンピュータ等を用いることにより、小型で経済的なシステムとして実現することができる。

【００１３】

【発明が解決しようとする課題】かかる音声・顔動画像出力装置をパーソナルコンピュータ上で実現させる場合、処理量削減のため、上述したように合成画像を予め作成しておいてそれらの画像を入力された文章に応じて切り換えて表示することが一般に行われている。これらの装置において合成音声と顔動画像を生成するにあたっては、声質、画面上での画像の表示場所、表示倍率などのパラメータは、表示するシステムに初期値として予め設定されたもの（パラメータ入力部７で予め入力されたもの）が使われる。

【００１４】このように従来の装置では合成音声の声質と顔動画像の生成態様を受信側で予め設定しておくものであるが、それら予め登録されてある顔画像の人物と声質が例えばメッセージに対して釣り合っていないような場合、それをみる人に不自然な感じを与えてしまうことになる。

【００１５】また、この装置を電子メールなどに用いた場合などに代表されるように、文章情報と合成画像を作った人とその文章情報を実際に音声と動画像で表示して見る人とが異なる場合、文章情報と合成画像を作った人が希望するような声質や画像の大きさで、受信側において発音・画像表示されるとは限らず、この結果、送り側の人の意図とは全く違う印象を受信側の人に与えてしまう可能性がある。

【００１６】つまり従来の装置では、音声と動画像でメッセージを伝えるにあたっての声質や顔の容貌などから表示側の人が受ける印象は表示側で予め設定したパラメータによって決まってしまうことになり、情報の作成側の人が意図した印象表現を表示側の人に的確に伝えることができなかった。

【００１７】本発明はかかる問題点に鑑みてなされたものであり、その目的とするところは、表示側で文章情報に基づいて合成音声あるいは顔合成画像を表示するにあたり、その文章等の作成側の人が意図した通りの表示が可能となるようにすることにある。

【００１８】

【課題を解決するための手段】図１は本発明に係る原理説明図である。本発明においては、一つの形態として、任意の文章情報からそれに対応する合成音声および該合成音声に合わせて口が動く人物の顔の合成動画像を生成する合成画像表示システムにおいて、文章情報の作成側において顔の合成画像を作成する際に、表示側における合成音声と合成動画像の生成態様を決めるための各種パラメータを合成画像に付加して表示側に渡すように構成されたことを特徴とする合成画像表示システムが提供される。

【００１９】上記の各種パラメータは合成音声の声質、合成動画像を表示する際の表示倍率、表示位置を含むパラメータとすることができる。

【００２０】また本発明においては、他の形態として、受信した伝送データを合成動画像データ、文章情報、各種パラメータに分離する伝送データ入力手段と、文章情報に基づいて合成音声を生成し出力する音声合成手段と、伝送データ入力手段で分離された合成動画像データをファイリングする画像メモリと、文章情報をその文章情報を発声したときの一連の口形の動きを表す口形符号の系列に変換する変換手段と、文章情報に基づいて該音声合成手段から出力される合成音声の各音節の発音時間を計算して各音声の切れ目のタイミングを推定する発音時間計算手段と、発音時間計算手段で推定した各音節の切れ目のタイミングで表示画像を該変換手段からの口形符号に対応した口形画像に切り換える制御を行う表示制御手段と、伝送データ入力手段で分離された各種パラメ

ータを対応する内部回路に送るパラメータ入力手段とを備えた合成画像表示システムが提供される。

【００２１】

【作用】本発明の合成画像表示方式においては、送信側において、表示に必要な顔画像を合成した際、その同じデータに表示側で合成した人が希望する合成画像の表示倍率や表示位置、合成音声の声質、その他のパラメータを埋め込む。表示側では、システムの初期値としてその画像データに埋め込まれた値を用いる。これにより、顔画像を合成した人の意図した通りに表示システム側で合成音声と合成画像を生成することができる。

【００２２】また本発明の他の形態の合成画像表示システムにおいては、伝送データ入力手段で受信した伝送データを合成動画像データ、文章情報、各種パラメータに分離し、文章分解手段で文章情報を解析して発音制御データを生成し、音声合成手段でこの発音制御データに基づいて合成音声を生成し出力し、受信した合成動画像データを画像メモリにファイリングし、変換手段で発音制御データを口形符号の系列に変換し、発音時間計算手段で発音制御データに基づいて音声合成手段で発音される各音節の発音時間をそれぞれ計算して各音節の切れ目のタイミングを推定し、画像表示制御手段で各音節のタイミング信号に合わせてその音節の口形画像を画像メモリから読み出すように制御し、受信した各種パラメータをパラメータ入力手段で対応する内部回路に送って合成音声と合成画像の生成態様を文章作成側の人が意図したものとなるようにする。

【００２３】

【実施例】以下、図面を参照して本発明の実施例を説明する。図２は本発明の一実施例としての合成画像表示システムによる音声・顔動画像出力装置が示される。図２において、文章分解部１、規則音声合成部２、音声／口形変換部３、発音時間計算部４、画像表示制御部５、画像メモリ６は前述の従来例で説明したものと同じものである。

【００２４】従来装置との相違点として、送信側から送られてきた伝送データには、本来の文章情報の他に、送信側の人が受信側で合成され表示されることを希望する顔画像と口形画像等の合成画像データ、さらにその人が希望する画面上での表示倍率、表示位置、声質、その他のパラメータが合成画像データに埋め込まれている。

【００２５】図３には送信側においてこれらのパラメータを伝送データに埋め込むための処理の概念が示される。受信側での表示を希望する顔画像の原画像に基づいて顔モデルへのマッピングを行い、各口形のパラメータを用いて合成画像データを作成し、これに表示倍率、表示位置、声質、その他の受け側に与える印象に係わるパラメータを埋め込む。これとは別に文章情報を作成し、双方を伝送データとして受信側に送る。この場合、パラメータが埋め込まれた合成画像データを一度送ってしまえば、後は文章情報を繰り返し送るだけでよい。

【００２６】受信側ではこの伝送データは伝送データ入力部８に入力され、ここで、文章情報、合成画像データ、各種パラメータに分離され、文章情報は文章分解部１に、合成画像データは画像メモリ６に、各種パラメータはパラメータ入力部７にそれぞれ送られる。

【００２７】パラメータ入力部７はこの各種パラメータを受け取ると、この各種パラメータを調べて、音声合成に関する声質等のパラメータは規則音声合成部２に、画像の表示倍率、表示位置等の画像に関するパラメータは画像表示制御部５と画像メモリ６にそれぞれ送る。

【００２８】このように構成することで、受信側では、表示すべき顔画像とシステムの初期値として埋め込むパラメータとして、送信側の人が希望した顔画像と、表示倍率、表示位置、声質、その他のパラメータを用いることができる。よって送信側の人の意図通りの音声と画像で受信側の表示システムにメッセージを表示させることができる。

【００２９】本発明の実施にあたっては種々の変形形態が可能である。例えば、上述の実施例では口形として７種類の画像を用いる場合について説明したが、もちろん本発明はこれに限られるものではなく、より自然に近い口の動きを合成するためにはこの口形の画像の種類をさらに増やしてもよい。また上述の実施例では表示側で合成する顔画像の動き部分として口領域の動きを取り上げたが、これに限られるものではなく、例えば口の動きに加えて、文章に合わせて目の動きなども変化させるようにすれば、より表情豊かなＡＶメッセージを受け側に送ることができる。

【００３０】また上述の実施例では本発明をＡＶ電子メールに適用した場合について説明したが、本発明はこれに限られるものではなく、音声・顔動画像出力装置単体に適用することも可能であるし、あるいは、例えば音声認識技術によりリアルタイムに発声音声の音素の認識が可能となれば、通常の電話をかけるだけで受信者側に話し手の顔の表情も動画像で表示できるという擬似テレビ電話等のサービスに適用することも可能である。

【００３１】

【発明の効果】以上に説明したように、本発明によれば、受信側で文章情報に基づいて合成音声あるいは顔合成画像を表示するにあたり、その文章の送り側の人が意図した通りの表示が可能となる。

【図面の簡単な説明】

【図１】本発明に係る原理説明図である。

【図２】本発明の一実施例としての合成画像表示システムによる音声・動画像出力装置を示す図である。

【図３】実施例システムによる送り側での処理概念を説明する図である。

【図４】従来の音声・動画像出力装置を示す図である。

【符号の説明】

7

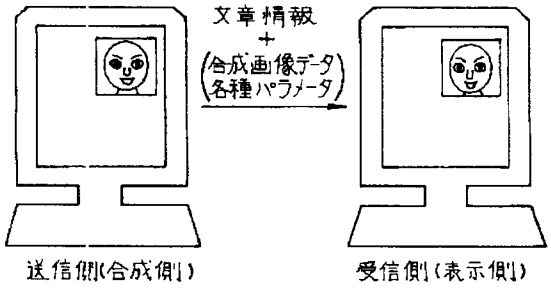
1　文章分解部
2　規則音声合成部
3　音声／口形変換部
4　発音時間計算部

8

＊5　画像表示制御部
6　画像メモリ
7　パラメータ入力部
＊　8　伝送データ入力部

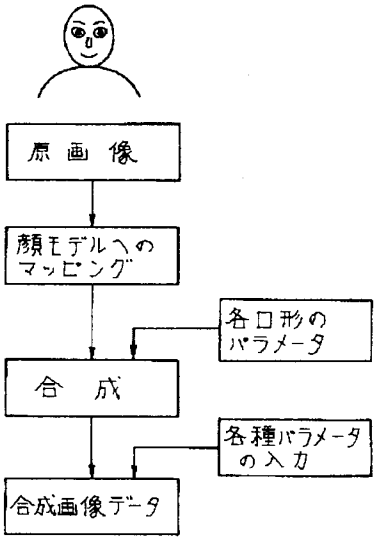
【図１】

本発明に係る原理説明図



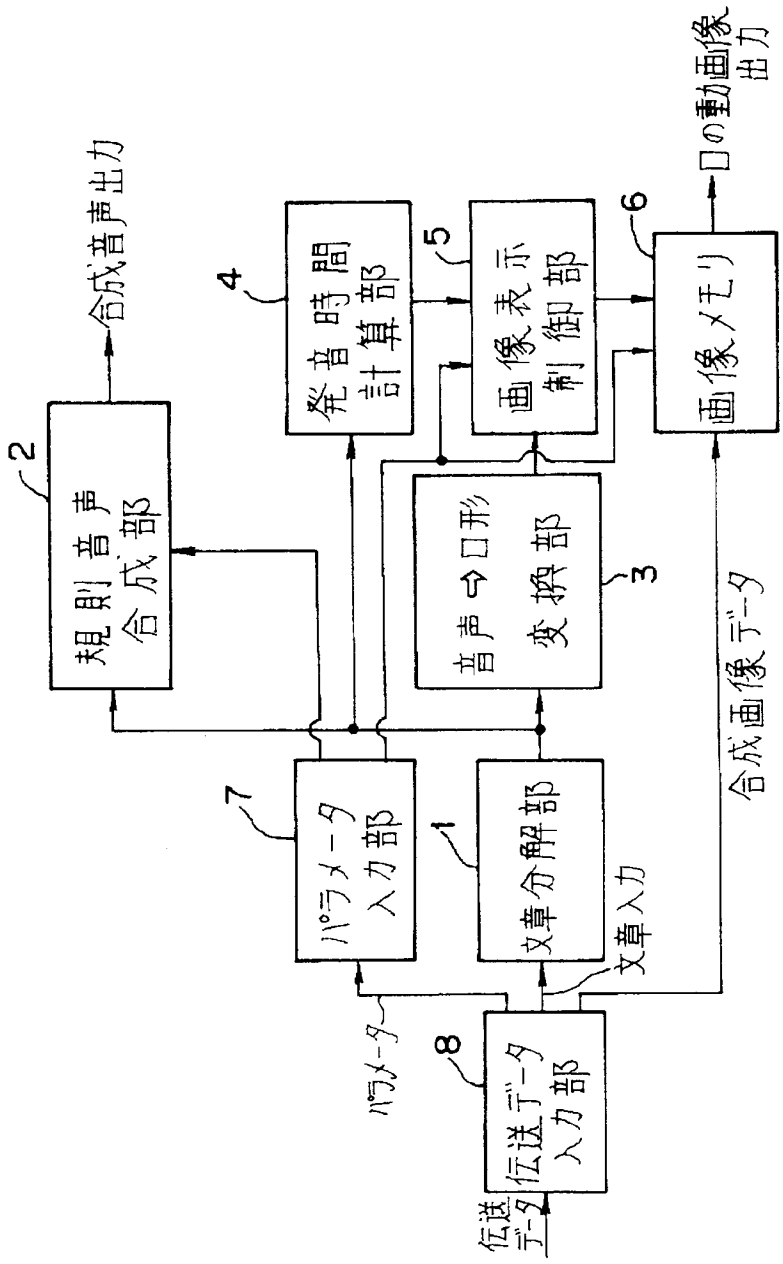
【図３】

送信側の処理



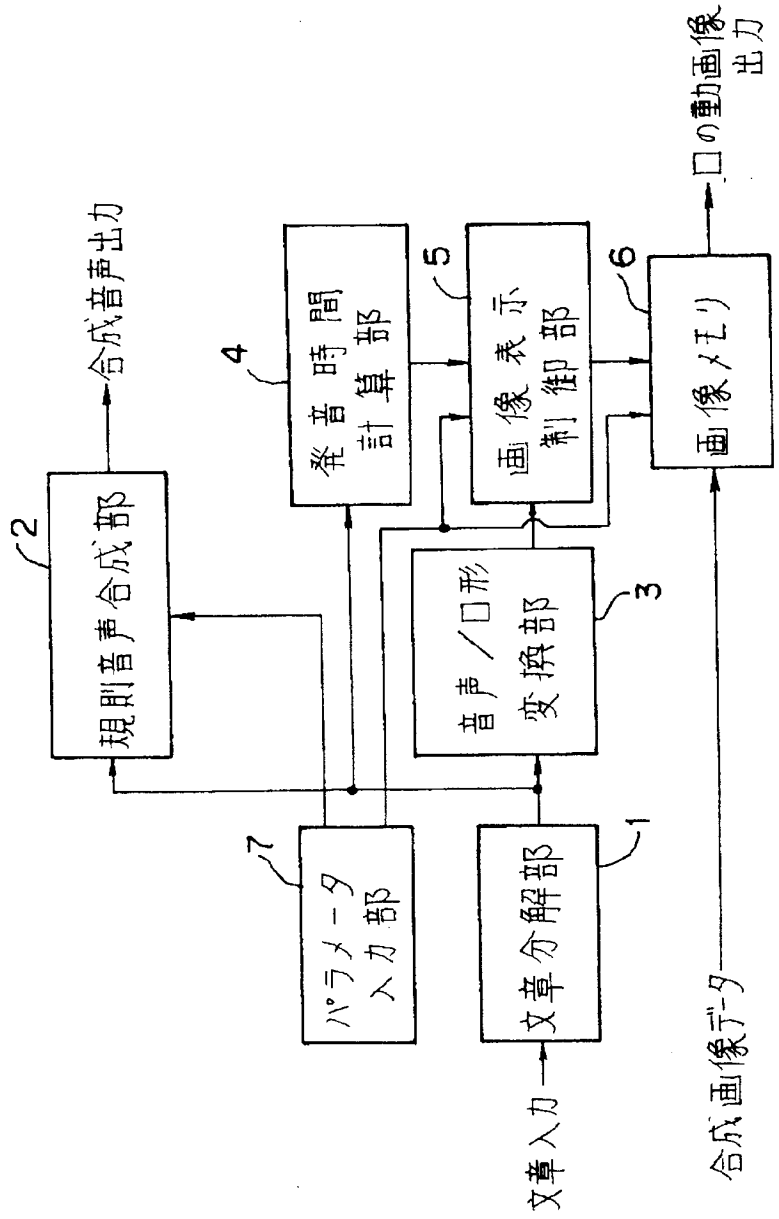
【図2】

本発明の実施例


合成音声出力
口の動画像出力
2
規則音声合成部
4
発音時間計算部
5
画像表示制御部
6
画像メモリ
音声⇔口形変換部
3
合成画像データ
7
パラメータ入力部
1
文章分解部
文章入力
パラメータ
8
伝送データ入力部
伝送データ

【図4】

従来例


口の動画像出力
合成音声出力
4
5
6
発音時間計算部
画像表示制御部
画像メモリ
2
規則音声合成部
音声／口形変換部
3
7
パラメータ入力部
文章分解部
1
文章入力
合成画像データ